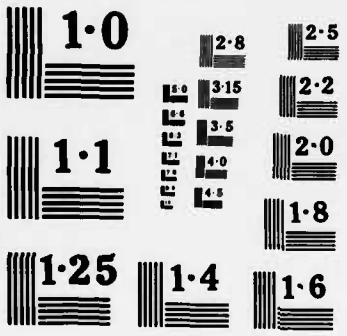
AD-A158 133 PENALIZED LIKELIHOOD FOR GENERAL SEMI-PARAMETRIC REGRESSIOO MODELS(U) WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CEETER P J GREEN MAY 85 MRC-TSR-2819 1/1

UNCLASSIFIED DAAG29-80-C-0041 F/G 12/1 NN



MRC Technical Summary Report #2819

PENALIZED LIKELIHOOD FOR GENERAL SEMI-PARAMETRIC REGRESSION MODELS

Peter J. Green

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

May 1985

(Received March 19, 1985)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

(A)

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

PENALIZED LIKELIHOOD FOR GENERAL SEMI-PARAMETRIC REGRESSION MODELS

Peter J. Green

Technical Summary Report #2819
May 1985

ABSTRACT

This paper examines maximum penalized likelihood estimation in the context of general regression problems, characterized as probability models with composite likelihood functions. The emphasis is on the common situation where a parametric model is considered satisfactory but for inhomogeneity with respect to a few extra variables. A finite-dimensional formulation is adopted, using a suitable set of basis functions. Appropriate definitions of deviance, degrees of freedom, and residual are provided, and the method of cross-validation for choice of the tuning constant is discussed. Quadratic approximations are derived for all the required statistics.

Additional keywords: algorithms; smoothing; goodness of fit tests; nonlinear regression.

AMS (MOS) Subject Classifications: 62F10, 62G05, 62J02, 65U05

Key Words: Basis functions, composite likelihood function, cross-validation, decoupled likelihood, maximum penalized likelihood estimation, nonlinear regression, roughness penalty, smoothing.

Work Unit Number 4 (Statistics and Probability)

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Statisticians and their clients have considerable experience constructing, fitting and interpreting parametric models for their data. But in many situations a completely parametric model is inappropriate. This often arises when a parametric relationship can be justified on grounds of theory or experience, but is suspected of varying slowly in time or space. The statistician is then reluctant either to abandon the parametric model, which is credible and useful, or to force a particular dependence on time or space into the model, without guidance on the form of this dependence. A semi-parametric approach is needed.

In this paper the method of maximum penalized likelihood is advocated for such problems, combining the ideas of fitting parameters by maximizing likelihood whilst smoothing with respect to the extraneous variables. This method is well-known in non-parametric linear regression, where it includes spline smoothing, and also in such problems as non-parametric density estimation. The present context is a rather general class of regression models, allowing nonlinearity, statistical dependence, and arbitrary error distributions.

A basic algorithm for fitting the model is derived, and asymptotic theory briefly dicussed. Various related statistics facilitating assessment of *goodness*-of-fit and determination of the appropriate degree of smoothing are constructed, together with quadratic approximations likely to make numerical computational economical.

PENALIZED LIKELIHOOD FOR GENERAL SEMI-PARAMETRIC REGRESSION MODELS

Peter J. Green

1. INTRODUCTION

It frequently arises that a statistician has some faith in the validity of a certain parametric statistical model for his data, but for some suspected inhomogeneity with respect to one or more extraneous variables. Typically, such variables might represent space or time, the relationship between them and the response is not of primary interest, and the statistician is inhibited from extending his parametric model to encompass them because of a lack of experience, information or theory about the form of their relationship. A simple example might arise with binomial data from different geographical locations where it might be quite reasonable to model the response probability as a logistic regression on various explanatory factors or covariates, but influenced also by environmental effects, unknown in form but believed to vary smoothly with location.

In such situations, procedures derived from penalized likelihoods (Good and Gaskins (1971), Silverman (1984)) may well be appropriate. The purpose of this paper is to examine properties of such methods in the context of the rather general class of regression models used by Green (1984), characterized as probability models expressed as composite likelihood functions. (This is not to claim that such a view of regression is universally appropriate). The methods discussed combine the ideas of fitting the parametric part of the model by maximizing likelihood whilst smoothing with respect to the extraneous variables.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

We consider a log-likelihood function $L(y;\psi)$ for the data y in term of a n-vector of predictors ψ. It will be helpful to allow the dimensionality of the data vector to be greater or less than n. The form of this probability model will not be seriously questioned, but rather the focus of attention will be on the dependence of ψ on explanatory factors and covariates, symbolised by x, and extraneous variables, denoted by t. We suppose that ψ has a prescribed functional form in terms of x,t, a p-vector of unknown parameters β, and an unknown real-valued function γ defined on the space in which t is measured (typically R^1 or R^2).

Thus the complete model is

$$L(y;\psi(x,\beta,t,\gamma)) = L(\psi(\beta,\gamma)) \ , \tag{1.1}$$

say, where the observed y,x and t may be omitted from the notation. Here only $\beta \in R^p$ and γ, lying in some prescribed linear space of functions G, are unknown and our principal interest is in β.

For an example, consider a logistic regression model in which the 'intercept' term varies in time. Then if the ith observation is of y_i successes out of m_i trials, with covariates $\{x_{ij}\}$ recorded at time t_i, we would write

$$L(\psi) = \sum_{i=1}^{n} \{y_i \log \psi_i + (m_i - y_i)\log(1 - \psi_i)\}$$

where

$$\psi_i = \{1 + \exp(-\sum x_{ij}\beta_j - \gamma(t_i))\}^{-1} \ .$$

This simple problem is typical of many where maximum likelihood leads to over-fitting, in the absence of any restriction on the form of the function γ. At least there will be unidentifiability of parameters; possibly a 'Dirac catastrophe'. In the context of density estimation, Good and Gaskins (1971)

proposed maximizing instead the penalized likelihood

$$P = L(\psi(\beta,\gamma)) - \frac{1}{2}\lambda J(\gamma) \qquad (1.2)$$

where J is a roughness penalty, increasing as the function γ becomes less smooth, and λ is a non-negative tuning constant or hyperparameter which may be adjusted to control the smoothness of the fitted γ. There is of course a Bayesian interpretation; see Section 4.

If the likelihood L is that of independent observations y_i, Normally distributed with means ψ_i linear in β and $\gamma(t_i)$, then this penalized likelihood approach is equivalent to a semi-parametric linear regression as proposed by Green, Jennison and Seheult (1983, 1985) in the context of agricultural field experiments, Engle, Granger, Rice and Weiss (1983), in an economic problem and Wahba (1984), who with co-workers has developed a considerable body of theory for such 'partial-spline' methods. Use of penalized likelihood in simple generalized linear models is discussed by O'Sullivan in his thesis (1983), by O'Sullivan, Yandell and Raynor (1984) and by Silverman (1985). Leonard (1982) considers such methods for a variety of curve estimation problems from a full-blooded empirical Bayesian perspective. In all of these papers the parametric part of the model (x,β) is not present, but Wahba (1985) remarks that the ideas of partial splines may be combined with penalized likelihood for generalized linear models. In none of these papers in the general regression model (1.1) addressed, and typically identification of γ is not regarded as of subsidiary importance to the efficient estimation of β.

2. THE ESTIMATION PROCEDURE

We begin by apparently compromising the generality of prescription of our problem. As it stands, (1.1) allows the predictor ψ to depend on infinitely many values of the function γ. In practice, therefore, discretization will be necessary at some stage. Following a suggestion of Leonard (1982), in maximizing (1.2), we will restrict γ to lie in a finite-dimensional subspace of G, namely $F = \text{span}\{\phi_j,\ j = 1,2,\ldots,q\}$, for a prescribed set of q basis functions. We write

$$\theta(\beta,\xi) = \psi(\beta, \sum_{j=1}^{q} \xi_j\phi_j) , \tag{2.1}$$

and will further restrict attention to roughness penalties of the form $J(\sum_{j=1}^{q} \xi_j\phi_j) = \xi^T K\xi$ for some fixed $q \times q$ non-negative definite matrix K.

It may seem that we are abandoning our intended semi-parametric framework, but it should be stressed that q, while it may be somewhat less than n, will still be 'large', and parametric estimation of ξ will not be appropriate. Further, the intention is that F and G should in practical terms be indistinguishable. This will entail appropriate choice of $\{\phi_j\}$ as, for example, a large class of orthogonal polynomials or trigonometric functions in t. This choice will depend on the observed values of t, and will also determine K. The precise quadratic form of the roughness penalty is hardly necessary in what follows, but it simplifies the algebra and is not likely to make any practical difference.

There may in fact be no restriction at all. In non-parametric regression, G is typically a reproducing kernel Hilbert space, on which J is a squared semi-norm. Suppose ψ depends only on $\{\gamma(t_i),\ i = 1,2,\ldots,q\}$. Then since

$$\min\{J(\gamma) : \gamma(t_i) = \xi_i,\ i = 1,2,\ldots,q\} = \xi^T K \xi$$

for a certain K, and we may choose F to consist of a basis of spline functions with $\phi_j(t_i) = \delta_{ij}$, the original and the restricted problem have the same solution so far as values of β and of $\{\gamma(t_i)\}$ are concerned.

We therefore maximize

$$P = L(\theta(\beta,\xi)) - \frac{1}{2}\lambda\xi^T K \xi \qquad (2.2)$$

over $\beta \in R^p$, $\xi \in R^q$, where θ is a prescribed R^n-valued function. This revised formulation has the further advantage of allowing certain new problems into our framework, that could not otherwise be naturally described with a vector of predictors ψ of finite length: see example (d) in the next section.

We will only be concerned with problems where likelihood methods are appropriate: we suppose sufficient regularity that L is approximately quadratic near the 'true values' β_o, ξ_o. A modification to an iteratively reweighted least squares algorithm derived from the Newton-Raphson method, with Fisher scoring, (see Green, 1984) should therefore be appropriate.

Write

$$u = \frac{\partial L}{\partial \theta}, \quad A = E\left[-\frac{\partial^2 L}{\partial \theta \partial \theta^T}\right], \quad D = \frac{\partial \theta}{\partial \beta}, \quad E = \frac{\partial \theta}{\partial \xi}.$$

The scores u form an n-vector, and the matrices A, D, and E are $n \times n$, $n \times p$ and $n \times q$. All of these quantities in general depend on β and ξ, (in the case of u and A only through θ), but these dependencies will be suppressed from the notation. The expectation is taken at the current values of β and ξ. Differentiating (2.2) gives the modified likelihood equations

$$D^T u = 0 \qquad (2.3)$$

$$E^T u = \lambda K \xi .$$

Their solution gives our required maximum penalized likelihood estimates (MPLE's) $\hat{\beta}$, $\hat{\xi}$. Typically these equations are nonlinear and require iterative solution. The Newton-Raphson method with expected second derivatives involves successively replacing trial estimates (β,ξ), at which u,A,D and E are evaluated, by (β^*,ξ^*) where

$$\begin{pmatrix} D^TAD & D^TAE \\ E^TAD & E^TAE + \lambda K \end{pmatrix}\begin{pmatrix} \beta^* - \beta \\ \xi^* - \xi \end{pmatrix} = \begin{pmatrix} D^Tu \\ E^Tu - \lambda K\xi \end{pmatrix}$$

or, equivalently,

$$G\begin{pmatrix} \beta^* \\ \xi^* \end{pmatrix} = \begin{pmatrix} D^TAY \\ E^TAY \end{pmatrix} \tag{2.4}$$

where

$$G = H + \begin{pmatrix} 0 & 0 \\ 0 & \lambda K \end{pmatrix}, \quad H = \begin{pmatrix} D^T \\ E^T \end{pmatrix}A[D : E] ,$$

and

$$Y = A^{-1}u + D\beta + E\xi . \tag{2.5}$$

These equations have the form of a combination of weighted normal equations, for β^*, and generalized ridge regression equations, for ξ^*. The two ingredients are found separately in Green (1984) and O'Sullivan, Yandell and Raynor (1984). See also Silverman (1985, Section 8.1).

We can now move towards stating conditions on the model (1.1), (2.1) for this approach to be applicable. First represent K as L^TL, where L is $r \times q$ of full rank r, which is usually less than q. If so, then K has a non-trivial null space: Let T be $q \times (q - r)$ such that $LT = 0$ and $[L^T : T]$ is non-singular. Our conditions are that for all β,ξ, the matrix A is non-singular, and D,E and $[D : ET]$ have full rank p,q and $p + q - r$ respectively. We may then proceed with any positive finite λ: the matrix G is non-singular. Convergence of the iteration (2.4) is

not guaranteed, but in practice will usually occur rather rapidly for sensible initial values. The algorithm has at least a fixed-point justification: if (2.4) gives $\beta^* = \beta$ and $\xi^* = \xi$, then (2.3) is satisfied.

Jointly with Dr. Brian Yandell, the author is developing various implementations of the basic algorithm (2.4, 2.5). Details will appear elsewhere.

3. SPECIAL CASES

The general model (1.1) makes no assumptions about the independence of random terms, or additivity and linearity among systematic components. Of course such simplifications are sometimes available. If the log-likelihood L is that of n independent observations $\{y_i\}$ each indexed by the corresponding ψ_i, then A is diagonal. If θ is linear in β or ξ, then D or E will be constant. Such properties may be exploited in algorithms, but do not affect a general treatment.

(a) The Linear Normal Case

If the observations y are independently Normally distributed, $y \sim N(\theta,\sigma^2 I)$, with a linear parameterization $\theta = D\beta + E\xi$, then D and E are constant, and $A = \sigma^{-2}I$. The scale factor σ^2 factorizes from both sides of (2.4), so may be ignored : this is an example of a more general phenomenon : see Section 9. The artificial response Y in (2.5) is identically y, and no iteration is necessary. If E,ξ,λ and K are omitted, we have the ordinary linear model. If D and β are omitted instead then we have a model including ridge regression (when $K = I$, we obtain $\hat{\xi} = (E^TE + \lambda I)^{-1}E^Ty$) and spline smoothing (as described in Section 2). With both D and $E = I$ present, this covers the least-squares

smoothing approach to the analysis of agricultural field trials due to Green, Jennison and Seheult (1983, 1985). They used a roughness penalty based on differencing ξ from neighbouring plots, for example

$$\xi^T K \xi = \sum_i (\xi_i - 2\xi_{i+1} + \xi_{i+2})^2 \qquad (3.1)$$

In this application, D represents a designed experiment, and the resulting methodology may be related to other more classical approaches (see Green, 1985).

In all these special cases, it may be more natural to focus on least-squares rather than Normal theory/maximum likelihood as the basic principle.

(b) Logistic Regression.

To continue the example from Section 1, with now $\theta_i = \{1 + \exp(-\sum x_{ij}\beta_j - \xi_i)\}^{-1}$; we have $u_i = (y_i - m_i\theta_i)/\{\theta_i(1 - \theta_i)\}$, A is diagonal with $A_{ii} = m_i/\{\theta_i(1 - \theta_i)\}$, E is diagonal with $E_{ii} = \theta_i(1 - \theta_i)$, and $D_{ij} = \theta_i(1 - \theta_i)x_{ij}$. The equations (2.4) are no longer fixed and iteration is necessary. An appropriate form for K will depend on the temporal or spatial configuration of the $\{t_i\}$: see Section 4.

(c) A Grouped Continuous Model.

For non-parametric regression of ordered categorical data on a single explanatory variable, the following model may be appropriate. For $r = 1,2,\ldots,R$ we have a S-vector multinomial response $\{y_{rs}, s = 1,\ldots,s\}$ with associated probabilities $\{p_{rs}\}$ assumed to satisfy

$$\sum_{i=1}^{s} p_{ri} = \Psi(\beta_s - \xi_r)$$

for some prescribed distribution function Ψ, where $\xi_r = \gamma(t_r)$ and t_r is the value of the explanatory variable for this response. This grouped continuous model is equivalent to the assumption of a latent continuous

variable with distribution function $\Psi(\cdot - \xi_r)$ which is categorized into S classes at the unknown cutpoints $\{\beta_1,\beta_2,\ldots,\beta_{s-1}\}$ to yield the observed frequencies $\{y_{rs}\}$. See McCullagh (1980) for a complete discussion. This falls into our present framework if we take θ as

$\{\theta_{rs} = \beta_s - \xi_r;\ r = 1,2,\ldots,R;\ s = 1,2,\ldots,S-1\}$, so that $p_{rs} = \Psi(\theta_{rs}) - \Psi(\theta_{r\,s-1})$. The matrix A is no longer diagonal, but D and E have a very simple form. For identifiability, one component of ξ must be held fixed and omitted from equations (2.4).

(d) Multiple Inhomogeneous Poisson Processes

Suppose m point processes are observed: the i^{th} process is observed for the time interval $(s_i,t_i]$, and yields events $\{y_{ik},\ k = 1,2,\ldots,n_i\}$. The observation intervals may depend on the realization: for example, each process may be observed until the first event, as in survival analysis. If the processes may be modelled as independent inhomogeneous Poisson processes with rates $\psi\left(\sum_{j=1}^{p} x_{ij}\beta_j,\gamma(t)\right)$ at time t for the i^{th} process, involving an unknown linear combination of covariates $\{x_{ij}\}$, the appropriate log-likelihood is

$$L = \sum_{i=1}^{m}\sum_{k=1}^{n_i} \log \psi\left(\sum x_{ij}\beta_j,\gamma(y_{ik})\right) - \sum_{i=1}^{m}\int_{s_i}^{t_i} \psi\left(\sum x_{ij}\beta_j,\gamma(y)\right)dy\ .$$

Here, ψ is prescribed, but β and the function γ are to be estimated. Again, a Dirac catastrophe prevents maximum likelihood estimation of β without restrictions on γ. Under the assumption that the rates are simply $\exp\left(\sum x_{ij}\beta_j + \gamma(t)\right)$ with appropriate stopping rules, this is the proportional hazards model (Cox, 1972, Anderson and Senthilselvan, 1980), and estimation of β is possible via partial likelihood. Otherwise, the approach described in Section 2 may be necessary.

4. CHOICE OF ROUGHNESS PENALTY.

Various authors have remarked in different contexts that choice of the amount of smoothing (λ in equation (2.2)) is more important than the smoothing kernel K itself. This might be amended by adding that the null space of the roughness, the space $\{\xi : \xi^T K \xi = 0\}$ may also be important; for vectors in this null space, since they are not penalized in (2.2), are implicitly also fitted as covariates.

Spline smoothers would choose a penalty of the form

$$J(\gamma) = \int_a^b (\gamma^{(m)}(t))^2 dt \qquad (4.1)$$

for a curve on a single-dimensional variable. As mentioned in Section 2 this is equivalent to our approach; the kernel K is given by

$$K_{ij} = \int_a^b \phi_i^{(m)}(t)\phi_j^{(m)}(t)dt \quad i,j = 1,2,\ldots,q .$$

the rank of K will be $q - m$ for any spline basis $\{\phi_k : k = 1,\ldots,q\}$, the null space of K consisting of those ξ for which $\sum_{k=1}^{q} \xi_k \phi_k$ is a polynomial of degree $(m - 1)$ or less.

Wahba (1978) derives spline smoothing from a Bayesian model in which an appropriate prior, which is partially improper, is constructed on a space of smooth functions; see also Silverman (1985). In the notation above, the prior is a multivariate Normal distribution for ξ with mean 0 and inverse variance matrix λK. Impropriety of the prior is equivalent to deficient rank in K. In our present partially parametric context, with an uninformative prior for β as an additional ingredient, the maximum penalized likelihood estimate for (β,ξ) is the mode of the corresponding posterior

distribution. We can make the prior more explicit, whilst avoiding impropriety, as follows. Let L and T be as constructed in Section 2; then we can generate the prior for ξ as

$$\xi = T\delta + L^T(LL^T)^{-1}\varepsilon \qquad (4.2)$$

where δ is a fixed $(q - r)$-vector and ε an r-vector of zero-mean, uncorrelated Normally distributed random variables with variance λ^{-1}. We can see that the penalty term $\lambda\xi^T K\xi$ is indeed twice the negative of the appropriate log-likelihood term.

Leonard (1982) provides a more completely Bayesian approach for the non-parametric case, again using a Gaussian process as a prior for γ: specifically he recommends an Ornstein-Uhlenbeck process, with two hyperparameters in place of λ, for the difference between the derivative of γ and a prescribed or estimated base curve. The full empirical Bayesian approach allows estimation of the hyperparameters.

If the observations are located at equally-spaced $\{t_i\}$ on a line, the squared m^{th} derivative penalty (4.1) will in practice be indistinguishable from that involving m^{th} differences, for example (3.1), with a basis such that $\phi_j(t_i) = \delta_{ij}$. Use of other roughness penalties was also considered by Green et al. (1985).

When, and only when, the log-likelihood L is that of a Normal distribution with expectation θ linear in ξ, addition of the roughness penalty corresponds to use of a 'random effects' model for ξ, or equivalently, so far as β is concerned, to modification of the assumed variance structure for y. See Green (1985).

Whatever form of K is chosen, the tuning constant λ controls the relative impact of roughness, as judged by K, and error, determined by the likelihood. The extreme cases $\lambda \to \infty$ and 0, between which we wish to

compromise, can be interpreted as follows. As $\lambda \to \infty$, the likelihood is maximized subject to a roughness penalty of 0, that is we restrict to the purely parametric model $L(\theta(\beta, T\delta))$. As $\lambda \to 0$, the problem degenerates to the minimization of $\xi^T K \xi$ subject to the 'interpolation' condition that $\theta(\beta,\xi)$ maximizes $L(\theta(\beta,\xi))$: whether this is a constraint on ξ depends on the form of θ.

5. ASYMPTOTICS AND STANDARD ERRORS

For various reasons, a rigorous treatment of the asymptotic theory of the maximum penalized likelihood estimates $(\hat{\beta},\hat{\xi})$ is extremely difficult: the arbitrariness of the probability model, and its parameterization, the presence of the roughness penalty, and flexibility over which parameter dimensions the asymptotics are to be with respect to. We introduce a hyperparameter N, upon which q and n may depend, and consider the limit $N \to \infty$. The dimension p of the parameter β will remain fixed, but other quantities such as λ and K may implicitly depend on N. It is clear that we do not require $n \to \infty$: for example in parametric logistic regression the standard asymptotic theory applies if either $n \to \infty$ or $\min\{m_i\} \to \infty$. For the purely parametric case, a rather general result is given by McCullagh (1983), which we believe might be extended to the present case if q remains fixed as $N \to \infty$. His condition is essentially that $N^{-1} D^T A D$ has a non-singular limit.

In the spline smoothing case the usual asymptotic framework is one where the locations $\{t_i\}$ of the observations become increasingly dense in a finite interval as $N \to \infty$ (Craven and Wahba, 1979; Cox, 1984). This is true also for the non-Normal fully parametric case: see O'Sullivan (1983) and Cox and O'Sullivan (1983). However for the partial-spline/least-squares smoothing

case it is intuitively plausible that estimates of at least certain contrasts of β should be consistent and asymptotically Normal even when the spacings between the $\{t_i\}$ do not decrease.

Here we will only attempt crude heuristics. Suppose that the model $L(\theta(\beta,\xi))$ is correct, so that there are true values β_o, ξ_o. (This is restrictive: we would hope to estimate β_o satisfactorily without such strong assumptions on the non-parametric part of the model). We will use the symbols $\hat{}$, $_o$ etc., on other quantities, so that for example $\hat{u}$ means $u(\theta(\hat{\beta},\hat{\xi}))$. For some point $(\bar{\beta},\bar{\xi})$ between $(\hat{\beta},\hat{\xi})$ and (β_o,ξ_o), we have by Taylor's theorem:

$$\begin{pmatrix} \frac{\partial P}{\partial \beta} \\ \frac{\partial P}{\partial \xi} \end{pmatrix}_{\hat{\beta},\hat{\xi}} = \begin{pmatrix} \frac{\partial P}{\partial \beta} \\ \frac{\partial P}{\partial \xi} \end{pmatrix}_{\beta_o,\xi_o} + \left[\frac{\partial^2 P}{\partial \begin{pmatrix} \beta \\ \xi \end{pmatrix} \begin{pmatrix} \beta \\ \xi \end{pmatrix}^T} \right]_{\bar{\beta},\bar{\xi}} \begin{pmatrix} \hat{\beta} - \beta_o \\ \hat{\xi} - \xi_o \end{pmatrix} .$$

The left-hand side is 0, so

$$\begin{pmatrix} \hat{\beta} - \beta_o \\ \hat{\xi} - \xi_o \end{pmatrix} \simeq \bar{G}^{-1} \begin{pmatrix} D_o^T u_o \\ E_o^T u_o - \lambda K \xi_o \end{pmatrix} .$$

where here, and later, we use $\simeq$ to mean asymptotically equal, under unspecified assumptions. We have $E(u_o) = 0$, $\text{var}(u_o) = A_o$, and $\bar{G} \simeq G_o \simeq \hat{G}$, $A_o \simeq \hat{A}$, $D_o \simeq \hat{D}$, $E_o \simeq \hat{E}$. So if the central limit theorem applies in the usual way, we would expect that the asymptotic distribution of $(\hat{\beta},\hat{\xi})$ is approximately

$$\begin{pmatrix} \hat{\beta} \\ \hat{\xi} \end{pmatrix} \sim N\left[\begin{pmatrix} \beta_o \\ \xi_o \end{pmatrix} - G^{-1} \begin{pmatrix} 0 \\ \lambda K \xi_o \end{pmatrix}, G^{-1} H G^{-1} \right] \qquad (5.1)$$

where G and H may be estimated by $\hat{G}$ and $\hat{H}$.

Obviously much remains to be proved, but these arguments do suggest that we may at least use (5.1) to derive nominal standard errors for $\hat{\beta}$, assuming that we have chosen λ depending on N in such a way that the penalized likelihood estimates are consistent.

6. DEVIANCE AND DEGREES OF FREEDOM.

In the spirit of the previous section, an analogue of the usual likelihood ratio statistic may be constructed, under a regularity condition. We need the notion of 'saturated model' to be well-defined, in the following way. Let $L_{max} = \sup_{\theta} L(\theta) = L(\theta')$, say, be the maximum value attained by $L(\theta)$ when the n-vector θ of predictors is freed from its functional dependence on β and ξ. We suppose $L_{max} < \infty$. Define the deviance associated with a particular regression function $\theta(\beta,\xi)$ and given values β and ξ by

$$\Delta = 2\{L_{max} - L(\theta(\beta,\xi))\} ;$$

note that because of the penalization, our estimates do not minimize Δ, but rather the penalized deviance

$$\Delta + \lambda\xi^T K\xi = 2\{L_{max} - P(\beta,\xi)\} .$$

In certain circumstances the information matrix A may be approximately constant, and the likelihood approximately quadratic, near θ' and $\theta(\hat{\beta},\hat{\xi})$, in which case since $u(\theta') = 0$ we have $\hat{u} = u(\theta(\hat{\beta},\hat{\xi})) \simeq \hat{A}(\theta' - \hat{\theta})$, whence $\Delta \simeq \hat{u}^T\hat{A}^{-1}\hat{u}$. We will refer to this latter expression as the linearized deviance.

Just as is standard practice with generalized linear models (Nelder and Wedderburn, 1972), model selection in the form of choice of a regression function $\theta(\beta,\xi)$ may be based on nominal significance tests using the approximate asymptotic distribution of the deviance. Similar heuristics to

those used above justify a χ^2 approximation on an appropriate "equivalent" degrees of freedom, which will not in general be integral, defined via the asymptotic expectation of the deviance.

We have

$$E(\Delta) \simeq E(\hat{u}^T\hat{A}^{-1}\hat{u}) .$$

But $\hat{A} \simeq A_o$ and $\hat{u} \simeq u_0 - A_o(D_o(\hat{\beta} - \beta_o) + E_o(\hat{\xi} - \xi_o))$, so from the asymptotic distributions of $(\hat{\beta},\hat{\xi})$ in Section 5, we find

$$E(\hat{u}^TA^{-1}\hat{u}) \simeq (0 \vdots \lambda K\xi_0)^TG^{-1}\begin{pmatrix} D^T \\ E^T \end{pmatrix}AA^{-1}A(D \vdots E)G^{-1}\begin{pmatrix} 0 \\ \lambda K\xi_0 \end{pmatrix}$$

$$+ \operatorname{tr}\{A^{-1}[(I - A(D \vdots E)G^{-1}\begin{pmatrix} D^T \\ E^T \end{pmatrix})A(I - (D \vdots E)G^{-1}\begin{pmatrix} D^T \\ E^T \end{pmatrix}A)]\}$$

$$= \operatorname{tr}(M^2) + Q_1 - Q_2 \quad \text{say}$$

where $M = \left(I - B^T(D \vdots E)G^{-1}\begin{pmatrix} D^T \\ E^T \end{pmatrix}B\right)$, B is a square root of $A = BB^T$, and $Q_j = \xi_o^T\{\lambda K(E^TAE + \lambda K - E^TAD(D^TAD)^{-1}D^TAE)^{-1}\}^j\lambda K\xi_o$. Note that $0 \leq Q_j \leq Q_{j+1}$, with mutual equality if and only if $K\xi_o = 0$. Turning now to the penalty term evaluated at the MPL estimates,

$$E(\lambda\hat{\xi}^TK\hat{\xi}) = [\xi_o^T + (0 \vdots (-\lambda K\xi_o)^T)G^{-1}\begin{pmatrix} 0 \\ I \end{pmatrix}]\lambda K[\xi_o + (0 \vdots I)G^{-1}\begin{pmatrix} 0 \\ -\lambda K\xi_o \end{pmatrix}]$$

$$+ \operatorname{tr}(\lambda K(0 \vdots I)G^{-1}\begin{pmatrix} D^T \\ E^T \end{pmatrix}A(D \vdots E)G^{-1}\begin{pmatrix} 0 \\ I \end{pmatrix})$$

$$= \operatorname{tr}(M - M^2) + Q_0 - 2Q_1 + Q_2 .$$

Thus the penalized deviance has asymptotic expectation

$$E(\Delta + \lambda\hat{\xi}^TK\hat{\xi}) \simeq \operatorname{tr}(M) + (Q_0 - Q_1) .$$

The utility of these expressions is limited by presence of the correction terms involving quadratic forms in the unknown true ξ_0. Three possible

approaches include (i) neglecting these terms, leading to an underestimate of degrees of freedom and hence conservative tests, (ii) estimating these terms from the data, a necessarily hazardous activity, or (iii) replacing them by their expectations under the prior distribution (4.2) that was used to justify the penalty function.

For the latter approach, we have $E(\lambda\xi_o^T K\xi_o) = E(\lambda\varepsilon^T\varepsilon) = r$, and for any $q \times q$ matrix N, $E((\lambda K\xi_o)^T N(\lambda K\xi_o)) = E(\lambda L^T\varepsilon)^T N(\lambda L^T\varepsilon))$ $= tr(\lambda KN)$. Thus $E(Q_j) = t_j$, say, where $t_0 = r$, and $t_j = tr\{\lambda K(E^TAE + \lambda K - E^TAD(D^TAD)^{-1}D^TAE)\}^j$ for $j = 1,2,3,\ldots$. But a little manipulation reveals that $tr(M^j) = n - (p + q) + t_j$ for $j = 1,2,3,\ldots$, so that under the prior (4.2) the terms in the penalty function have asymptotic expectations, at $(\hat{\beta},\hat{\xi})$, with the simple forms:

$$E(\hat{u}^TA^{-1}\hat{u}) \simeq n - (p + q) + t_1 = tr(M) = \nu ,$$

say

$$E(\lambda\hat{\xi}^TK\hat{\xi}) \simeq t_0 - t_1 = n - (p + q) + r - \nu$$

It may be shown that for any λ, ν satisfies the inequality

$$n - rk[D \vdots E] \leq \nu \leq n - (p + q) + r .$$

(Its exact form as a function of λ when D,E,A and K are constant is given in Green (1985). In the notation of that paper, R is ET and V is $I + \lambda^{-1}E(E^TE)^{-1}L^T(L(E^TE)^{-1}L^T)^{-2}L(E^TE)^{-1}E^T$). Combining this with the information above about asymptotic expectations supports the use of ν as a surrogate error degrees-of-freedom. Parameters corresponding to the columns of $[D \vdots ET]$ (i.e. β and δ) are always fitted, requiring $(p + q - r)$ degrees of freedom, and the remaining $n - (p + q) + r - \nu$ are associated with the non-parametric component of ξ permitted by $\lambda < \infty$. Of course, no formal distribution theory involving ν is known: it should only be used informally to gauge model adequacy as measured by the deviance.

7. CROSS-VALIDATION

Model selection by means of cross-validation was discussed in a systematic way by Stone (1974). Its use in determining an appropriate degree of smoothing in non-parametric regression problems has been enthusiastically espoused by Wahba and co-workers in the past ten years, and, in a refined form known as generalized cross-validation (GCV) (Wahba, 1977) seems to have become the de facto standard approach. In the linear problem, GCV has additional invariance over the ordinary version, it has now become computationally practicable, and it is known to provide an asympotically optimal degree of smoothing in a predictive mean-square sense. O'Sullivan (1983) generalizes the application of GCV to generalized linear models by transcribing a formula from the linear case, without deriving the criterion afresh from its plausible first principles. Thus we attempt to do here, for our more general class of regression problems.

The basic idea in cross-validation is to delete one observation at a time from the data set, and endeavour to predict it from the model as fitted to the remaining observations. The smoothing parameter is chosen to optimize the overall quality of prediction. The appropriate generalization of this 'delete-one' operation in our model $L(\theta(\beta,\xi))$ consists of <u>decoupling</u> each component of θ in turn from its dependence on β and ξ. The predictive discrepancy will be measured in likelihood or deviance terms.

The decoupling is achieved by the introduction of dummy covariates. For some genrality, let F be an arbitrary $n \times f$ matrix, $f \geq 1$, and let $\tilde{\beta}$, $\tilde{\xi}$ and $\tilde{\tau}$ maximize the penalized decoupled likelihood $L(\theta(\tilde{\beta},\tilde{\xi}) + F\tilde{\tau}) - \frac{1}{2}\tilde{\xi}^T K\tilde{\xi}$. We define the predictive discrepancy in the column space of F as the non-negative quantity

$$\Delta^{\uparrow}(F) = 2\{L(\theta(\tilde{\beta},\tilde{\xi}) + F\tilde{\tau}) - L(\theta(\tilde{\beta},\tilde{\xi}))\} \; ; \qquad (7.1)$$

if this is zero then the decoupled estimates $(\tilde{\beta},\tilde{\xi})$ coincide with $(\hat{\beta},\hat{\xi})$. We will average $\Delta^{\dagger}(F)$ over an appropriate set of directions to give an overall predictive discrepancy, but first we obtain a linear approximation for $\Delta^{\dagger}(F)$.

At $(\tilde{\beta},\tilde{\xi},\tilde{\tau})$ we have

$$D^T\tilde{u} = 0$$

$$E^T\tilde{u} = \lambda K\tilde{\xi}$$

$$F^T\tilde{u} = 0$$

where

$$\tilde{u} = u(\theta(\tilde{\beta},\tilde{\xi}) + F\tilde{\tau})$$

$$\simeq u(\theta(\hat{\beta},\hat{\xi})) - A(D(\tilde{\beta} - \hat{\beta}) + E(\tilde{\xi} - \hat{\xi}) + F\tilde{\tau}) \ .$$

But we know that $D^T\hat{u} = 0$ and $E^T\hat{u} = \lambda K\hat{\xi}$, so by subtraction, treating A,D and E as fixed (evaluated at $(\hat{\beta},\hat{\xi})$, say), we have

$$\begin{pmatrix} D^TAD & D^TAE & D^TAF \\ E^TAD & E^TAE + \lambda K & E^TAF \\ F^TAD & F^TAE & F^TAF \end{pmatrix} \begin{pmatrix} \tilde{\beta} - \hat{\beta} \\ \tilde{\xi} - \hat{\xi} \\ \tilde{\tau} \end{pmatrix} \simeq \begin{pmatrix} 0 \\ 0 \\ F^T\hat{u} \end{pmatrix}$$

whence $\tilde{\beta}$ and $\tilde{\xi}$ may be eliminated to give

$$\tilde{\tau} \simeq \left(F^TAF - F^TA(D \vdots E)G^{-1}\begin{pmatrix} D^T \\ E^T \end{pmatrix}AF\right)^{-1}F^T\hat{u} = (F^TBMB^TF)^{-1}F^T\hat{u} \ .$$

So, by linearizing the expression (8.1) and noting that $(\tilde{\beta},\tilde{\xi},\tilde{\gamma})$ maximizes the first term, penalized, we have

$$\Delta^{\dagger}(F) \simeq (F\tilde{\tau})^TA(F\tilde{\tau}) = \hat{u}^TF(F^TBMB^TF)^{-1}F^TAF(F^TBMB^TF)^{-1}F^T\hat{u} \ . \qquad (7.2)$$

This expression measures the result of decoupling any number f of components of θ; for an analogue of delete-one cross-validation we set $f = 1$. For

example if $F = e^{(i)}$, the unit vector in the i^{th} coordinate direction,

$$\Delta^{\dagger}(e^{(i)}) \simeq \frac{A_{ii}\hat{u}_i^2}{\{(BMB^T)_{ii}\}^2} .$$

If A is not diagonal, we may prefer to rotate R^n and have

$$\Delta^{\dagger}((B^{-1})^T e^{(i)}) \simeq (B^{-1}\hat{u})_i^2 / M_{ii}^2 .$$

In generalized cross-validation the individual predictive discrepancies are combined over different directions by a weighted sun enjoying certain invariance properties. Let $w_i = M_{ii}/tr(M)$, so that $\sum w_i = 1$, and define the GCV criterion

$$V(\lambda) = \sum_{i=1}^{n} w_i^2 \Delta^{\dagger}((B^{-1})^T e^{(i)}) \simeq (\hat{u}^T A^{-1}\hat{u})/(tr(M))^2 \simeq \Delta/\nu^2 .$$

We can choose λ to minimize this quantity, which has the same form as that used by O'Sullivan (1983).

That this is the correct weighting of the individual predictive discrepancies can be seen by examining the invariance properties of $V(\lambda)$. Re-parameterization by invertible appropriately differentiable transformations of θ,β and ξ does not change the model; it alters u,A,D and E, but $\hat{u}^T A^{-1}\hat{u}$, Δ, M and ν remain invariant, and so therefore does $V(\lambda)$.

One justification for use of the GCV criterion in the linear (spline) case is provided by the result of Craven and Wahba (1979) stating that such a criterion is asymptotically optimal in the sense of minimizing the mean-squared error $R(\lambda) = \frac{1}{n}\sum_{i=1}^{n} (\hat{\gamma}(t_i) - \gamma_o(t_i))^2$. (Of course, this property may be shared by many other criteria for choosing λ). The only natural expression of $R(\lambda)$ in likelihood terms seems to be via the divergence defined by Kullback and Leibler (1951). We define $R(\lambda)$ so that

$$nR(\lambda) = 2E_{\theta_o}(L(\theta_o) - L(\hat{\theta})) \simeq (\hat{\theta} - \theta_o)^T A(\hat{\theta} - \theta_o) \simeq (\hat{u} - u_o)^T A^{-1}(\hat{u} - u_o)$$

to give an appropriate weighted m.s.e. for $(\hat{\beta},\hat{\xi})$, which makes connections with the linearized deviance apparent. Cross-validation and Kullback-Leibler distance are also discussed by Bowman, Hall and Titterington (1984). O'Sullivan (1983) sketches an argument suggesting that the Craven and Wahba result extends to the generalized linear model case, with a definition of $R(\lambda)$ equivalent to the above; it therefore seems likely that if his proof could be rigorized, it might apply to the present more general setup as well.

Note that by arguments similar to those of Section 6, we have

$$E(R(\lambda)) \simeq Q_1 - Q_2 + tr((I - M)^2) ,$$

whose expectation under the prior for ξ_o is just $tr(I - M)$.

8. RESIDUALS.

How best to define residuals depends very much on the purpose to which they are to be put. The multitude of definitions available even in simple linear regression models (see Cook and Weisberg, 1982) strongly suggests that even more alternatives will be available in our present general context. Here we attempt only a limited discussion.

We seek residuals primarily for diagnostic purposes, and, in view of our reliance on the likelihood function $L(\theta(\beta,\xi))$, prefer these to be likelihood-based and associated with the predictors θ rather than directly with the observations. Use of such residuals for diagnosis of data-inadequacy will require inspection of the likelihood function to determine the data-points instrumental in giving a particular component of θ a large residual. Detection of model inadequacy can proceed more directly, and note that we not desire invariance of residuals to transformation of θ itself.

The likelihood emphasis suggests concentrating on the deviance Δ, defined in Section 6. Restricting attention to one or more particular components of θ, define

$$\Delta(F) = 2\{\sup_{\tau} L(\theta(\hat{\beta},\hat{\xi}) + F\tau) - L(\theta(\hat{\beta},\hat{\xi}))\} , \qquad (8.1)$$

twice the maximum increase in log-likelihood attained by freeing θ from its dependence on β and ξ, in the directions spanned by columns of F. If F is non-singular, $\Delta(F) = \Delta$. Note that $\Delta(F) \leq \Delta^{\dagger}(F)$; the sole difference between the two quantities lies in the inclusion or exclusion of the corresponding components of θ from the fitting of the model. Also note that the maximum penalized-likelihood ratio statistic $2\{\sup P(\theta(\beta,\xi) + F\tau) - \sup P(\theta(\beta,\xi))\}$ lies between $\Delta(F)$ and $\Delta^{+}(F)$.

Choosing a single coordinate direction $e^{(i)}$ for F, we obtain the raw deviance and discrepancy $\Delta(e^{(i)})$ and $\Delta^{\dagger}(e^{(i)})$ respectively, which we abbreviate as Δ_i and $\Delta_i^{\dagger}$. The raw deviances Δ_i have been customarily used to define residuals in generalized linear models (see discussion in Green, 1984). Finally, we denote the signed square-roots by $z_i = \text{sign}(\tau_{max})\sqrt{\Delta_i}$ and $z_i^{\dagger} = \text{sign}(\tau_{max})\sqrt{\Delta_i^{\dagger}}$, where τ_{max} denotes the value of τ in the maximization of (8.1) and in (7.1) respectively.

These concepts tie in well with other treatments of residuals. In the case of Normal linear regression (with known variances = 1, say, for simplicity), z_i and $z_i^{\dagger}$ are just the ordinary and predicted residuals, respectively, of Cook and Weisberg (1982, Chapter 2). These are known to be correlated, and improperly standardized for variance. Cook and Weisberg point out that z_i and $z_i^{\dagger}$ respectively under- and over-emphasize discrepancies for high-leverage data points. This difficulty will persist in the more general cases.

When y is distributed Normally with expectation $\theta = D\beta$ and known non-diagonal variance matrix V we find

$$z_i = \{V^{-1}(y - \theta)\}_i \{(V^{-1})_{ii}\}^{-1/2}$$

and

$$z_i^\dagger = \{V^{-1}(y - \theta)\}_i \{(V^{-1})_{ii}\}^{1/2} \{(V^{-1} - V^{-1}D(D^TD)^{-1}D^TV^{-1})_{ii}\}^{-1}$$

which are in fact y_i standardized by its expectation and variance assuming the parameters to be equal to their estimates with and without y_i, respectively, conditional on the other components of y.

In contrast to Jørgensen (1984), we believe that the use of unconditional moments in this standardization is inappropriate for dependent observations.

For these examples, the linearization leading to (8.2), and by a simpler argument to $\Delta(F) \simeq \hat{u}^T F(F^TAF)^{-1}F^T\hat{u}$ involves no approximation. In general, when the likelihood is not quadratic, replacing Δ_i and $\Delta_i^\dagger$ by these approximations leads to different residuals z_i and $z_i^\dagger$. The former have been called 'score residuals' (Jorgensen, 1983). But as pointed out by Green (1984), these score residuals are not appropriate when the quadratic approximation is badly wrong: for example, they are not monotonic functions of the observations in linear regression with a prescribed error density that is not log-concave.

9. NUISANCE PARAMETERS.

The approach to penalized likelihood estimation described here can handle with no difficulty certain types of nuisance parameter entering the probability model in addition to the predictors θ. Suppose, following Jørgensen (1983) that

$$L = L(y;\theta,\kappa) = c(y;\kappa) + \sigma(\kappa)t(y;\theta)$$

where $\sigma(\ )$, which we might term the precision parameter, is a scalar function of the possibly vector-valued nuisance parameter κ. See also Green (1984). This is in a sense the ultimate generalization of the property of generalized linear models in which the scale parameter factors out from the fitting procedure and is estimated at convergence from the deviance. Examples include the variance in the Normal distribution, the index in the gamma distribution, and also the extra parameter often allowed as a modification of the binomial or Poisson distributions to allow for 'over-dispersion'.

The maximum penalized likelihood estimates of β and ξ now satisfy

$$\sigma D^T\hat{u} = 0$$

$$\sigma E^T\hat{u} = \lambda K\hat{\xi} \ .$$

Fisher scoring is no longer available necessarily, because the expectation of $t(y;\theta)$ will in general involve κ, but if we write $A = -\dfrac{\partial^2 t}{\partial\theta\theta^T}$ then neglecting the second derivatives of θ with respect to β and ξ, (the "linearization method" of Jørgensen (1983)) we obtain the approximate Newton-Raphson iteration

$$\begin{pmatrix} D^TAD & D^TAE \\ E^TAD & E^TAE + \sigma^{-1}\lambda K \end{pmatrix}\begin{pmatrix} \beta^* \\ \xi^* \end{pmatrix} = \begin{pmatrix} D^TAY \\ E^TAY \end{pmatrix},$$

demonstrating that β and ξ can be estimated without paying any attention to the nuisance parameter κ, except that the value of the tuning constant λ is now effectively measured with respect to the unknown precision σ, a consequence that is not likely to be of any serious concern.

Acknowledgements: I am indebted to Tom Leonard for useful discussions on a Bayesian interpretation of the decoupling in Sections 7 and 8, and to Brian Yandell for suggestions on computing methods.

REFERENCES

Anderson, J. A. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. J. R. Statist. Soc. B, 42, 322-327.

Bowman, A. W., Hall, P., and Titterington, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. Biometrika, 71, 341-351.

Cook, R. D. and Weisberg, S. (1982). Residuals and influence in regression. Chapman and Hall, New York.

Cox, D. D. and O'Sullivan, F. (1983). Asymptotic analysis of the roots of penalized likelihood equations. Technical Report, Dept. of Statistics, University of Wisconsin-Madison.

Cox, D. D. (1984). Multivariate Smoothing Spline functions. SIAM J. Numer. Anal., 21, 789-813.

Cox, D. R. (1972). Regression models and life tables (with Discussion). J. R. Statist. Soc. B, 34, 187-202.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. 31, 377-403.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1983). Non-parametric estimates of the relation between weather and electricity demand. Discussion paper 83-17, Dept. of Economics, Univ. of California, San Diego.

Good, I. J. and Gaskins, R. A. (1971). Non-parametric roughness penalties for probability densities. Biometrika 58, 255-277.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). J. R. Statist. Soc. B, 46, 149-192.

_____ (1985). Linear models for field trials, smoothing and cross-validation. To appear in Biometrika.

Green, P. J., Jennison, C. and Seheult, A. H. (1983). Contribution to discussion of paper by G. N. Wilkinson, et al., J. R. Statist. Soc. B, 45, 193-195.

_____ (1985). Analysis of field experiments by least squares smoothing. J. R. Statist. Soc. B, 47.

Jørgensen, B. (1983). Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. Biometrika, 70, 19-28.

_____ (1984). Contribution to discussion of Green (1984).

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. Annals of Statistics, 22, 79-86.

Leonard, T. (1982). An empirical Bayesian approach to the smooth estimation of unknown functions. Tech. Summary Report 2339, MRC, University of Wisconsin-Madison.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). J. R. Statist. Soc. B, 42, 109-142.

_____ (1983). Quasi-likelihood functions. Ann. Stat. 11, 59-67.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. J. R. Statist. Soc. A, 135, 370-384.

O'Sullivan, F. (1983). The analysis of some penalized likelihood schemes. Ph.D. Thesis, Technical Report #726, Department of Statistics, University of Wisconsin-Madison.

O'Sullivan, F., Yandell, B. S. and Raynor, W. J. (1984). Automatic smoothing of regression functions in generalized linear models. Technical Report #734, Dept. of Statistics, University of Wisconsin-Madison.

Silverman, B. W. (1984). Penalized maximum likelihood estimation. To appear in Encyclopedia of Statistical Sciences.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). J. R. Statist. Soc. B, 47.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). J. R. Statist. Soc. B, 36, 111-147.

Wahba, G. (1977). A survey of some smoothing problems, and the method of generalized cross-validation for solving them. In Applications of Statistics (P. R. Krishnauah, ed.) 507-523. Amsterdam: North Holland.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. J. R. Statist. Soc. B, 40, 364, 372.

Wahba, G. (1984). Partial spline models for the semi-parametric estimation of functions of several variables, in Statistical Analysis of Time Series, Tokyo: Institute of Statistical Mathematics, 319-329.

Wahba, G. (1985). Comments to Projection Pursuit, by P. J. Huber, To appear in Ann. Statist.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2819	2. GOVT ACCESSION NO. AD-A158135	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PENALIZED LIKELIHOOD FOR GENERAL SEMI-PARAMETRIC REGRESSION MODELS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Peter J. Green		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE May 1985 13. NUMBER OF PAGES 26
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Basis functions, composite likelihood function, cross-validation, decoupled likelihood, maximum penalized likelihood estimation, nonlinear regression, roughness penalty, smoothing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper examines penalized likelihood estimation in the context of general regression problems, characterized as probability models with composite likelihood functions. The emphasis is on the common situation where a parametric model is considered satisfactory but for inhomogeneity with respect to a few extra variables. A finite-dimensional formulation is (cont.)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (cont.)

adopted, using a suitable set of basis functions. Appropriate definitions of deviance, degrees of freedom, and residual are provided, and the method of cross-validation for choice of the tuning constant is discussed. Quadratic approximations are derived for all the required statistics.

END

FILMED

10-85

DTIC

probability densities. Biometrika 58, 255-277.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). J. R. Statist. Soc. B, 46, 149-192.

of Wisconsin-Madison.

O'Sullivan, F., Yandell, B. S. and Raynor, W. J. (1984). Automatic smoothing of regression functions in generalized linear models. Technical Report #734, Dept. of Statistics, University of Wisconsin-Madison.

This paper examines penalized likelihood estimation in the context of general regression problems, characterized as probability models with composite likelihood functions. The emphasis is on the common situation where a parametric model is considered satisfactory but for inhomogeneity with respect to a few extra variables. A finite-dimensional formulation is (cont.)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)